New Search Paradigms

Innovation at the MIT Libraries
Search What?

*All the Libraries holdings*

- Books that we hold, or that other libraries hold if we can get them (e.g. via ILL, borrow direct)

- Journals, whether in print on the shelf or online at the publisher

- Archival and special collections, digitized or not

- And all that other stuff: images, music, other audio, video, maps, GIS datasets, statistical datasets, etc.

- And everything that we license online, or that is useful on the public Web
Search How?

- Fielded search
  - Author, title, subject date
  - Qualified by date, format, genre, etc.
  - Requires standardization or normalization

- Keyword search
  - Of full-text or metadata
  - Doesn’t work for non-textual materials

- Browse
  - Show users what “values” are available for every index
  - This is just another form of search
Search How?

- “Broadcast” search (aka federated search)
  - Many targets searched on-the-fly
  - Easy to add more targets (e.g. new licenses resources)
  - Hard to scale up to all the resources available

- Union catalog approach
  - Aggregate all searchable data locally
  - Think of Google, Worldcat, Proquest
Search Innovation Beyond MIT

Many libraries are experimenting with search

- NCSU
  new OPAC that supports faceted browsing over the MARC metadata

- University of Pennsylvania
  broadcast search of local databases (catalog, e-resources, IR, digital library collections, research guides, web pages, library FAQ and staff pages)

- California Digital Library
  “Common Framework” to integrate Metalib (via X-server), local digitized collections, online archival collections, new OPAC search tools, and more
Search Innovation Beyond MIT

- RLG’s RedLightGreen naïve user search system
- USC’s Ghandara project for integrated library e-collection searching
- JStor’s faceted browser
- University of Rochester’s eXtended Catalog project
- NLM’s Entrez PubMed

- And the list goes on and on…
Search Innovation Beyond MIT

- Beyond Libraries there are publishers (traditional and e-only), aggregators, campus/national/international e-learning and e-research systems (e.g. NPG’s Connotea, Atom Publishing Protocol, Elsevier’s Scopus, Thomson’s Web Citation Index, OKI)

- And then there’s Google, Google Scholar, Yahoo, CiteSeer, CiteULike, Amazon/A9, Blinx, YouTube, Flickr, Del.icio.us, Topicseek

Search is Hot!
MIT Libraries’ Response

What can we do right now?

What will we do in the future?
SIMILE Project

The Problem

DSpace is built on Dublin Core descriptive metadata

*But that isn’t good enough*

- We need to extend, qualify what’s there, invent whole new schemas for what isn’t there (e.g. policy metadata)
- We need RDF (Resource Description Framework)
SIMILE Project

If you have a lot of heterogeneous metadata in RDF, then what can you do?

Not fielded searching (don’t know what fields are there)

Plain keyword searching would lose the structure of the data, produce weird results

Faceted browsing!
Longwell

Longwell is a web-based RDF-powered highly-configurable faceted browser.

Can be customized for different purposes

- **Piggy-Bank** (Firefox extension to search/browse RDF stored on your personal computer)
- **Semantic Banks** (subject-based RDF collections on a server with a faceted browser web UI, ability to add user tags)
  
  Used for conferences (papers, presentations, people, schedules, etc), for projects, for content collections, etc.

More tools in the works!
DWell

- Another customization of *Longwell*, designed for DSpace sites

- Easy to integrate, immediate benefit to users
The End

Questions? Comments?